NBA Player Performance vs Salary: Final Project Report

Author: Godwin Omowele
Date: July 2025

---

1. Introduction & Problem Statement
NBA teams spend hundreds of millions on player contracts each season.
Goal: Measure how on-court performance relates to salary, and identify over- and under-paid players.
Business Impact:
- Data-driven contract negotiations
- Informed free-agency targeting
- Smarter draft and salary-cap management

---

2. Data & Wrangling
Sources:
- `data/nba_player_stats_checked.csv`
- `data/advanced_player_stats_checked.csv`
- `data/nba_salary_checked.csv`

Wrangling Steps:
1. Loaded all three CSVs and merged on `Player` (inner join).
2. Stripped “\$” and commas from `Salary`, cast to `float`.
3. Dropped identifier/rank columns (`Player`, `Team`, `Rk_x`, `Rk_y`, etc.).
4. Exported clean master file to `data/merged_stats.csv`.

---

3. Preprocessing
- **Dummy features:** One-hot encode all categorical columns.
- **Scaling:** Median-impute numeric missings, then standardize with `StandardScaler`.
- **Train/Test split:** 75% train / 25% test, `random_state=42`.

---

4. Modeling & Metrics

Model	RMSE	MAE	R^2
Linear Regression	2.010932e+16	3.992034e+15	-3.497084e+18
Random Forest	5 395 378.00	3 149 207.00	0.7482582
XGBoost	5 109 531.00	3 014 224.00	0.7742262

Final model: XGBoost (lowest RMSE, highest R

5. Residual Analysis
Residuals for the chosen model are roughly centered on zero with no obvious pattern, indicating a good fit.

---

6. Conclusions & Next Steps
- **Key insight:** Performance and salary correlate ($R^2 \approx$ <best_R2>), but there is still a \$[X]M “value variance” per player.
- **Recommendations:**
  1. Identify the top 10 under-paid players (predicted salary > actual salary) and target them in negotiations.
  2. Expand to multi-year analyses for trend detection.
- **Limitations:** Single-season snapshot; excludes bonuses and injury data.
- **Future work:** Add playoff/injury metrics, explore alternative model families, and test ensemble stacking.